Hierarchical Downlink Resource Management Framework for OFDMA based WiMAX Systems

Wang, Hua; Iversen, Villy Bæk

Published in:
Proceeding of IEEE WCNC 2008

Link to article, DOI:
10.1109/WCNC.2008.305

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the WCNC 2008 proceedings.

Hierarchical Downlink Resource Management Framework for OFDMA based WiMAX Systems

Hua Wang and Villy B. Iversen
Department of Communications, Optics & Materials
Technical University of Denmark, Lyngby, Denmark
Email: {huw, vbi}@com.dtu.dk

***Abstract*— IEEE 802.16, known as WiMAX, has received much attention recently for its capability to support multiple types of applications with diverse QoS requirements. Beyond what the standard has defined, radio resource management (RRM) still remains an open issue. In this paper, we propose a hierarchical downlink resource management framework for OFDMA based WiMAX systems. Our framework consists of a dynamic resource allocation (DRA) module and a connection admission control (CAC) module. DRA emphasizes on how to share the limited radio resources in term of subchannels and time slots among WiMAX subscribers belonging to different service classes with the objective of increasing the spectral efficiency while satisfying the diverse QoS requirements in each service class. CAC highlights how to limit the number of ongoing connections preventing the system capacity from being overused. Through system-level simulation, it is shown that the proposed framework can work adaptively and efficiently to improve the system performance in terms of high spectral efficiency and low outage probability.**

I. Introduction

Wireless metropolitan area network (WMAN) technology based on IEEE 802.16 standard and its evolutions has been developed to deliver a variety of multimedia services with different Quality-of-Service (QoS) requirements, such as throughput, delay, delay jitter, fairness and packet loss rate. Also known as Worldwide Interoperability for Microwave Access (WiMAX), IEEE 802.16 based technology is a promising alternative for last mile broadband wireless access. The physical layer specifications and MAC signaling protocols are well defined in the standard [9], however, radio resource management (RRM) (i.e., scheduling and admission control) still remains an open issue, which plays an important role in QoS provisioning for different types of services (i.e., real-time and non-real-time polling services require strict delay and throughput guarantees).

Orthogonal Frequency Division Multiple Access (OFDMA) is a physical layer specification for IEEE 802.16 systems. OFDMA builds on orthogonal frequency division multiplexing (OFDM), which is immune to intersymbol interference and frequency selective fading, as it divides the frequency band into a group of mutually orthogonal subcarriers, each having a much lower bandwidth than the coherence bandwidth of the channel. In multi-user environment, OFDMA enables dynamic assignment of subcarriers to different users at different time instances, to take advantage of the fact that at any time instance channel responses are different for different users at different subcarriers [4]. Thus, dynamic subcarrier assignment (DSA) and adaptive power allocation (APA) to multiple users can be employed to improve the system performance significantly.

Considerable amount of work has been done to investigate adaptive subcarrier and power allocation in OFDMA systems. In [1], the author addressed the problem of minimizing the total transmitted power at a given bit rate per terminal. Subcarrier and bit allocation were done dynamically through the use of nonlinear optimization with integer variables. The alternative objective of maximizing the overall system throughput under fairness constraints on users' data rates with upper bounded transmitted power has been considered in [2]. Such algorithms are often refereed to as *loading algorithms*.

In most loading algorithms, the QoS requirement of each user is usually defined in terms of a fixed data rate per frame. However, in practical communication systems, it is neither sufficient nor efficient to represent the QoS requirement by a fixed data rate per frame. The resource allocation problem for supporting both real-time (RT) and non-real-time (NRT) traffics in multimedia systems becomes much more complicated when diverse QoS requirements have to be considered. The transmission of RT packets can be delayed as long as the delay constraint is not violated, and the transmission of NRT packets can be more elastic. Therefore, efficient *packet-based scheduling algorithms* have been of interest.

Most of the packet-based scheduling algorithms can be categorized into one-level flat scheduler. In such approach, each connection is assigned a priority value based on some criterion and the connection with the highest priority is scheduled for transmission. However, due to different traffic patterns and diverse QoS requirements among rtPS, nrtPS and BE service classes, it is hard to well define a unified priority criterion. Thus, it is desirable to individually design the scheduling algorithm for each service class and separate the resource allocation from the packet scheduling [3].

In this paper, we propose a two-level hierarchical scheduler for the dynamic resource allocation (DRA) module and a measurement-based admission control strategy for the connection admission control (CAC) module. The DRA module is comprised of an aggregate resource allocator (ARA) and four class schedulers. For rtPS and nrtPS class schedulers, we introduce an extended Exponential Rule (EXP) algorithm, which tightly couples the packet scheduling and subcarrier allocation together to take advantage of the inter-dependencies between the PHY and MAC layers. For the aggregate resource

1525-3511/08/$25.00 ©2008 IEEE

allocator, an adaptive resource allocation scheme is presented. The proposed scheme first estimates the required amount of bandwidth in each class scheduler based on the backlogged traffic and the modulation efficiency. Then an exponentially smoothed curve with respect to QoS satisfaction is applied to adjust the estimated amount of bandwidth in order to increase the spectral efficiency while maintaining a guaranteed QoS performance. The proposed admission control strategy takes the current state of the network and class priority into considerations when admission decisions are made.

The rest of the paper is organized as follows. We first describe the considered system in the next section. In Section III, the proposed resource management framework is introduced, followed by the design of the class scheduler, the aggregate resource allocator, and admission control policies. Simulation environments and results are outlined and discussed in Section IV. Finally, a conclusion is drawn in Section V.

II. System Model

WiMAX technology supports both mesh and point-to-multipoint (PMP) networks. In this paper, we only investigate the WiMAX PMP network with OFDMA-TDD operation. We consider the downlink of a WiMAX system with U_s subcarriers and K users. The time axis is divided into frames. A frame is further divided into U_t time slots, each of which may contain one or several OFDM symbols. To reduce the resource addressing space, channel coherence in frequency and time is exploited by grouping I_s adjacent subcarriers and I_t time slots to form a basic resource unit (BRU). The size of BRU is adjusted so that the channel experiences flat fading in both frequency domain and time domain. Thus in each frame, there are $S = U_s/I_s$ subchannels in frequency and $N = U_t/I_t$ time slots in time, which corresponds to a total of $S*N$ BRUs. Each BRU can be assigned to different users and loaded with different power levels. In principle, adaptive power allocation can improve the system performance. However, some studies show that performance improvements are marginal over a wide range of SNRs due to the statistical effects [4]. Therefore, we assume that the total transmission power is equally distributed among all subchannels. Adaptive modulation and coding (AMC) scheme is applied to transmit data on each subchannel such that the highest possible rate is chosen. To utilize the PHY layer resources more efficiently, fragmentation at the MAC layer is enabled. For each connection, a separate queue with a maximum size of L PDUs is maintained, each of which is of fixed length with d information bits.

III. Downlink Resource Management Framework

The purpose of radio resource management is to increase the spectral efficiency while satisfying the diverse QoS requirements from different service classes. The proposed downlink resource management framework consists of a connection admission control (CAC) module and a dynamic resource allocation (DRA) module. CAC is responsible for preventing the system capacity from being overused by limiting the number of ongoing connections. DRA aims at an efficient usage of the scarce radio resources, while attaining certain fairness and QoS constraints among all the admitted users. Since it is difficult to formulate resource allocation as one mathematical optimization problem, the DRA module is further divided into two subproblems, i.e., a bandwidth distribution problem and a scheduling problem. Yet there is sufficient coupling between the allocator and the scheduler as the allocator is aware of the performance of the scheduler. An advantage of this two-level hierarchical resource allocation architecture is that the algorithms for the allocator and the scheduler can be modified independently of each other.

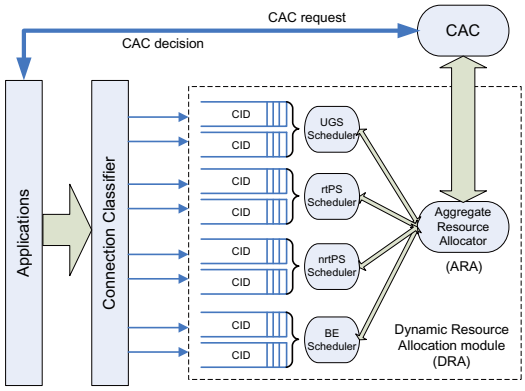

Fig. 1. Structure of the proposed downlink resource management framework for IEEE 802.16 systems

Fig. 1 depicts the proposed downlink resource management framework for IEEE 802.16 systems. When an application initiates a connection, it sends the connection request to the CAC module with connection type, traffic parameters and QoS requirements. Then the CAC module interacts with the DRA module to get the network state information and commits admission decisions. Arriving packets from the application layer are classified by the connection classifier according to their connection identifications (CID), and traffic types, and are sent to the corresponding service class and get queued. The DRA module is responsible for scheduling all the admitted connections. It consists of an aggregate resource allocator (ARA) and four class schedulers. The ARA distributes bandwidth to each class scheduler. Once the class scheduler receives bandwidth from the ARA, it schedules packets of its queues. In each class scheduler, because the incoming flows have similar traffic patterns and QoS requirements, the class scheduler has the freedom to independently choose its own scheduling algorithm which can best meet the QoS requirements. Therefore, this two-level hierarchical resource allocation module can have multiple scheduling criteria and better schedule packets in each service class than its one-level counterpart.

A. Class Scheduler Design

Class scheduler in each service class receives bandwidth from the ARA and involves in the allocation of subchannels, time slots and transmission powers among different users in its service queues. Scheduling algorithms designed for the class

schedulers should have the goal of maximizing the efficiency of resource utilization with satisfying QoS performance. Since the scheduling of UGS connections is well defined by the standard and BE connections do not have any specific QoS requirements, here we only focus on the design of rtPS and nrtPS class schedulers.

rtPS traffic is delay-sensitive and has strict delay requirement, while nrtPS traffic can tolerate longer delays, but requires a minimum throughput. The basic idea behind the proposed algorithm is that at each scheduling interval, if a PDU was scheduled for transmission on a specific subchannel, it is assigned a priority based on the instantaneous channel condition (PHY layer issue), as well as the QoS constraint (MAC layer issue). Then we can formulate the scheduling problem into a mathematical optimization problem with the objective of maximizing the total achievable priorities.

We apply an extended EXP algorithm as our priority function. It was proposed to provide QoS guarantees over a shared wireless link in terms of the average packet delay for RT traffic and a minimum throughput for NRT traffic [7].

For rtPS class scheduler, if the i^{th} PDU from user k's queue is scheduled for transmission on subchannel n, its priority is calculated as:

$$\mathbf{P}(k,i,n) = a_k \cdot \frac{\mu_{k,n}(t)}{\overline{\mu}_k(t)} \cdot \exp\Big(\frac{a_k W_{k,i}(t) - \overline{aW}}{1+\sqrt{\overline{aW}}}\Big) \quad (1)$$

where $\overline{aW} = \frac{1}{K}\sum_k a_k W_{k,1}(t)$, and $a_k = -\log\delta_k / T_{k,\max}$. $W_{k,i}(t)$ is the i^{th} PDU delay of user k at time t, $T_{k,\max}$ is the maximum allowable delay of user k, δ_k is the maximum outage probability of user k, $\mu_{k,n}(t)$ is the instantaneous channel rate with respect to the signal-to-noise ratio (SNR) and a predetermined target error probability if subchannel n is assigned to user k at time t, and $\overline{\mu}_k(t)$ is the exponential moving average (EMA) channel rate of user k with a smoothing factor t_c, calculated as:

$$\overline{\mu}_k(t) = (1-\frac{1}{t_c})\overline{\mu}_k(t-1) + \frac{1}{t_c}\mu_k(t) \quad (2)$$

where $\mu_k(t) = \sum_{n=1}^{N} c_{k,n} \cdot \mu_{k,n}(t)$ is the total channel rate of user k at time t. If subchannel n is assigned to user k, $c_{k,n} = 1$, otherwise $c_{k,n} = 0$.

For nrtPS class scheduler, the extended EXP algorithm is used in conjunction with a token bucket control to guarantee a minimum throughput [7]. We associate each NRT queue with a virtual token bucket. Tokens in each bucket arrive at a constant rate $r_{k,\text{req}}$, which is the required minimum throughput of user k. After a PDU is scheduled for service, the number of tokens in the corresponding token queue is reduced by the actual amount of data transmitted. The calculation of the priority for a nrtPS PDU is similar to Exp. (1), with the exception that $W_{k,i}(t)$ in nrtPS is defined as the virtual waiting time:

$$W_{k,i}(t) = \frac{\max\{0, V_k(t) - (i-1)\cdot d\}}{r_{k,\text{req}}} \quad k \in \text{nrtPS} \quad (3)$$

where $V_k(t)$ is the number of tokens associated with user k at time t, and d is the fixed MAC PDU size.

Let us define $\mathbf{u}(k,i,n)$ be the subchannel allocation indicator. That is, $\mathbf{u}(k,i,n) = 1$ means that the i^{th} PDU from user k is allocated on subchannel n for transmission, and $\mathbf{u}(k,i,n) = 0$ otherwise. Also let's define $\mathbf{m}(k,i,n)$ be the number of time slots needed on subchannel n if the i^{th} PDU from user k is scheduled for transmission on subchannel n, expressed as:

$$\mathbf{m}(k,i,n) = \lceil \frac{d}{\mu_{k,n}(t)} \rceil \quad (4)$$

where $\lceil x \rceil$ denotes the smallest integer larger than x.

Then, the scheduling problem can be mathematically formulated as follows:

$$\arg\max_{\mathbf{u}(k,i,n)} \sum_{k=1}^{K}\sum_{i=1}^{L}\sum_{n=1}^{N} \mathbf{u}(k,i,n)\cdot\mathbf{P}(k,i,n) \quad (5)$$

subject to:

$$\sum_{k=1}^{K}\sum_{i=1}^{L} \mathbf{u}(k,i,n)\cdot\mathbf{m}(k,i,n) \le S \quad \forall n \quad (6)$$

$$\sum_{n=1}^{N} \mathbf{u}(k,i,n) \le 1 \quad \forall k,i \quad (7)$$

$$\mathbf{u}(k,i,n) \in \{0,1\} \quad \forall k,i,n \quad (8)$$

The first constraint ensures that the allocated bandwidth do not exceed the total available bandwidth on each subchannel. The second constraint says that a PDU can only be transmitted via one subchannel. The instantaneous channel conditions and the QoS related parameters are embodied into the priority function $\mathbf{P}(k,i,n)$ with the objective of maximizing the total achievable priorities.

The above optimization problem can be solved by determining the values of integer variables $\mathbf{u}(k,i,n)$ through standard linear integer programming. The solution to the problem provides an optimal resource allocation. However, the computation complexity of the optimal solution is too high to be applied in practical systems. Instead, we have proposed a low complexity suboptimal heuristic algorithm in [8].

B. Aggregate Resource Allocator Design

The aggregate resource allocator (ARA) distributes the total available bandwidth among class schedulers. If the ARA does not allocate enough bandwidth to the class scheduler, the QoS requirements in the corresponding service class may not be guaranteed. On the other hand, if the ARA allocates too much bandwidth to the class scheduler, the allocated radio resource may not be utilized efficiently or even be wasted. Thus the resource allocation algorithm of ARA is a critical factor on the performance of the class scheduler and has to be carefully designed.

One possible solution is that the ARA distributes bandwidths among class schedulers in a static manner. That is to say, a pre-determined fixed amount of bandwidth is allocated to each class scheduler at every scheduling interval. This approach has the advantage of simplicity and works well when

the traffic pattern in each service class is regular and stable, which unfortunately is not always the case in data communications. Therefore, a dynamic resource allocation algorithm which can adapt to the traffic pattern and the performance of the class scheduler is believed to be a better solution.

In designing the proposed adaptive resource allocation algorithm, we have taken the following aspects into account: (i) the backlogged traffic; (ii) the modulation efficiency; (iii) the satisfaction of QoS requirement. The general idea is that the ARA first estimates the amount of bandwidths required in each class scheduler based on the backlogged traffic and the average modulation efficiency. Then the estimated bandwidths are further increased or decreased based on the QoS performance in each class scheduler.

We separate the bandwidth allocation of UGS class from the others as it has been defined by the standard. In UGS, the transmission mode at the PHY layer is fixed during the whole service time [9]. At the beginning of each frame, the ARA allocates a fixed amount of time slots $N_{\rm UGS} = \sum_{i\in\{\rm UGS\}} \theta_i$ to UGS connections based on their constant bit-rate requirements negotiated in the initial service access phase, where θ_i is the number of time slots required by UGS connection i. Let $N_{\rm total}$ be the total number of time slots in each frame, then the residual time slots after serving UGS class $N_{\rm rest} = N_{\rm total} - N_{\rm UGS}$ are distributed among rtPS, nrtPS and BE classes, which employ AMC scheme at the PHY layer.

For rtPS class, as each packet has a rigid delay requirement, the total sum of the current queue size in rtPS class is an appropriate measure for the backlogged traffic $B_{\rm rtPS}(t) = \sum_{i\in\{\rm rtPS\}} q_i(t)$, where $q_i(t)$ is the number of bits in queue i at time t. The average modulation efficiency $\overline{\mu}_{\rm rtPS}(t)$ is defined as the average number of bits carried per OFDM symbol over a sliding time window t_c. γ is a QoS related (i.e., maximum allowable delay in rtPS) parameter representing the proportion of backlogged traffic that has to be transmitted within each frame. Then the estimated number of time slots for rtPS class can be expressed as follows:

$$E_{\rm rtPS}(t) = \alpha(t) \cdot \frac{\gamma B_{\rm rtPS}(t)}{\overline{\mu}_{\rm rtPS}(t)} \tag{9}$$

where $\alpha(t)$ is a QoS-aware adjustment factor that is updated according to the performance of the class scheduler on a frame by frame basis. The basic idea is that when the class scheduler experiences good QoS satisfaction, the value of $\alpha(t)$ is decreased to save the bandwidth for other classes. Otherwise, the value of $\alpha(t)$ is increased to guarantee the required QoS. Towards this end, an exponentially smoothed curve is applied to adjust the value of $\alpha(t)$. The adjustment, which is $|\Delta\alpha(t)| = |\alpha(t) - \alpha(t-1)|$, is minor if the QoS outage probability is around a predefined threshold. Otherwise, $|\Delta\alpha(t)|$ is exponentially increased as either to increase or reduce the allocated bandwidth to the class scheduler. The calculation of $\Delta\alpha(t)$ is specified as follows:

$$\Delta\alpha(t) = \xi \cdot \frac{\exp(\beta \cdot d(t)) - 1}{\exp(\beta \cdot D_{\max}) - 1} \tag{10}$$

where $d(t)$ is further defined as:

$$d(t) = \begin{cases} \min\{\mathrm{P_r}(t) - T_{\rm h}, D_{\max}\} & \text{if } \mathrm{P_r}(t) \geq T_{\rm h} \\ \max\{\mathrm{P_r}(t) - T_{\rm h}, -D_{\max}\} & \text{if } \mathrm{P_r}(t) < T_{\rm h} \end{cases}$$

where $\mathrm{P_r}(t)$ is the delay outage probability at time t, $T_{\rm h}$ is the threshold of the outage probability, $D_{\max}$ is the truncated maximum value of $|d(t)|$, β is a shape factor which is used to tune the adaptation degree, and ξ is the maximum value of $|\Delta\alpha(t)|$. Term $(\exp(\beta \cdot d(t)) - 1)/(\exp(\beta \cdot D_{\max}) - 1)$ is a normalization function of $(\mathrm{P_r}(t) - T_{\rm h})$. When $\mathrm{P_r}(t)$ is close to $T_{\rm h}$, the normalized value is close to zero. Otherwise it increases exponentially to one. The bandwidth estimation procedure for rtPS class can be described as follows:

- **Step 1:** At each scheduling instant, calculate the backlogged traffic $B_{\rm rtPS}(t)$, the average modulation efficiency $\overline{\mu}_{\rm rtPS}(t)$, and the current delay outage probability $\mathrm{P_r}(t)$. Update the value of $\alpha(t)$:

$$\alpha(t) = \begin{cases} \min\{\alpha(t-1) + \Delta\alpha(t), \alpha_{\max}\} & \text{if } \mathrm{P_r}(t) \geq T_{\rm h} \\ \max\{\alpha(t-1) + \Delta\alpha(t), \alpha_{\min}\} & \text{if } \mathrm{P_r}(t) < T_{\rm h} \end{cases} \tag{11}$$

  where $\alpha_{\max}$ and $\alpha_{\min}$ are the maximum and minimum values of $\alpha(t)$, respectively.
- **Step 2:** Calculate the estimated bandwidth for rtPS class according to Exp. (9).

For nrtPS class, the bandwidth estimation procedure is the same as in rtPS class, except the definition of the backlogged traffic and the outage probability. Here we take the total number of virtual tokens associated with each queue as the measure for the backlogged traffic $B_{\rm nrtPS} = \sum_{i\in\{\rm nrtPS\}} v_i(t)$, where $v_i(t)$ is the number of virtual tokens in bucket i at time t. $\mathrm{P_r}(t)$ in nrtPS is the throughput outage probability.

At each scheduling instant, the ARA first allocates the amount of required bandwidth to UGS class ($N_{\rm UGS}$) and a minimum amount of bandwidth to BE class ($N_{\rm BE}^{\rm min}$). Once the ARA has estimated the amount of required bandwidth for rtPS and nrtPS classes (e.g., $E_{\rm rtPS}$ and $E_{\rm nrtPS}$), it checks the remaining bandwidth. If the remaining bandwidth is larger than the estimated sum of rtPS and nrtPS, the ARA first allocates $E_{\rm rtPS}$ and $E_{\rm nrtPS}$ to rtPS and nrtPS class schedulers respectively. Then the residual bandwidth is distributed among rtPS, nrtPS and BE classes proportional to their queue size $Q_{\rm rtPS}$, $Q_{\rm nrtPS}$, and $Q_{\rm BE}$. Otherwise, if the remaining bandwidth is smaller than the estimated sum of rtPS and nrtPS, the ARA distributes the bandwidth between rtPS and nrtPS classes proportional to their estimations $E_{\rm rtPS}$ and $E_{\rm nrtPS}$. It is worth mentioning that for the proportional fairness among class schedulers, each class scheduler is reserved a minimum amount of bandwidth. A detailed description of the proposed algorithm is listed in pseudocode 1.

C. Connection Admission Control

Connection admission control is a key component of QoS provisioning for wireless systems supporting multiple types of applications. It aims at maintaining the delivered QoS to

Algorithm 1 Adaptive bandwidth distribution algorithm in the Aggregate Resource Allocator (ARA)

Set initial N_{total} at the beginning of each frame
$N_{\text{UGS}} \leftarrow \sum_{i \in \{\text{UGS}\}} \theta_i$
$N_{\text{BE}} \leftarrow N_{\text{BE}}^{\min}$
$N_{\text{residual}} \leftarrow N_{\text{total}} - N_{\text{UGS}} - N_{\text{BE}}^{\min}$
Estimate the number of time slots allocated to rtPS class scheduler E_{rtPS} by Exp.(9)
Estimate the number of time slots allocated to nrtPS class scheduler E_{nrtPS} by Exp. (9)
if $N_{\text{residual}} \geq (E_{\text{rtPS}} + E_{\text{nrtPS}})$ **then**
  $N_{\text{residual}} \leftarrow N_{\text{residual}} - E_{\text{rtPS}} - E_{\text{nrtPS}}$
  $N_{\text{rtPS}} \leftarrow E_{\text{rtPS}} + N_{\text{residual}} \cdot \frac{Q_{\text{rtPS}}}{Q_{\text{rtPS}}+Q_{\text{nrtPS}}+Q_{\text{BE}}}$
  $N_{\text{nrtPS}} \leftarrow E_{\text{nrtPS}} + N_{\text{residual}} \cdot \frac{Q_{\text{nrtPS}}}{Q_{\text{rtPS}}+Q_{\text{nrtPS}}+Q_{\text{BE}}}$
  $N_{\text{BE}}^{'} \leftarrow N_{\text{residual}} \cdot \frac{Q_{\text{BE}}}{Q_{\text{rtPS}}+Q_{\text{nrtPS}}+Q_{\text{BE}}}$
  if $N_{\text{rtPS}} < N_{\text{rtPS}}^{\min}$ or $N_{\text{nrtPS}} < N_{\text{nrtPS}}^{\min}$ **then**
    Adjust the values of N_{rtPS}, N_{nrtPS} and $N_{\text{BE}}^{'}$ so that $N_{\text{rtPS}} \geq N_{\text{rtPS}}^{\min}$ and $N_{\text{nrtPS}} \geq N_{\text{nrtPS}}^{\min}$
  end if
else
  $N_{\text{rtPS}} \leftarrow N_{\text{rest}} \cdot \frac{E_{\text{rtPS}}}{E_{\text{rtPS}}+E_{\text{nrtPS}}}$
  $N_{\text{nrtPS}} \leftarrow N_{\text{rest}} \cdot \frac{E_{\text{nrtPS}}}{E_{\text{rtPS}}+E_{\text{nrtPS}}}$
  $N_{\text{BE}}^{'} \leftarrow 0$
  if $N_{\text{rtPS}} < N_{\text{rtPS}}^{\min}$ or $N_{\text{nrtPS}} < N_{\text{nrtPS}}^{\min}$ **then**
    Adjust the values of N_{rtPS} and N_{nrtPS} so that $N_{\text{rtPS}} \geq N_{\text{rtPS}}^{\min}$ and $N_{\text{nrtPS}} \geq N_{\text{nrtPS}}^{\min}$
  end if
end if
$N_{\text{BE}} \leftarrow N_{\text{BE}}^{'} + N_{\text{BE}}^{\min}$

different users at the target level by limiting the number of ongoing connections in the system. We propose a measurement-based approach as our CAC policy, of which a CAC decision is made based on the current state of the network. When a new connection is initiated, it informs the CAC module of the connection class (i.e., rtPS or nrtPS), the traffic parameters (i.e., arrival rate) and the QoS requirements (i.e., delay or minimum throughput). Then the CAC module estimates the required amount of bandwidth ΔN to accommodate the incoming connection and performs a CAC decision depending on the following conditions.

For UGS connections, as the transmission mode and the number of time slots allocated per frame are negotiated in the initial service access phase and are fixed during the whole service time, a simple threshold-based CAC is applied:

$$N_{\text{UGS}}^{\text{current}} + \Delta N_{\text{UGS}} < N_{\text{UGS}}^{\max} \tag{12}$$

where $N_{\text{UGS}}^{\text{current}}$ is the number of time slots occupied by ongoing UGS connections, and $N_{\text{UGS}}^{\max}$ is the maximum number of time slots that can be allocated to the UGS class scheduler. If this condition is satisfied, the incoming connection is accepted; otherwise, it is rejected.

For rtPS and nrtPS connections, when a new call arrives, the CAC module interacts with ARA in the DRA module and gets $\overline{E}_{\text{rtPS}}(t)$, $\overline{E}_{\text{nrtPS}}(t)$, which are the exponential moving average of the estimated bandwidth for rtPS and nrtPS classes respectively. If the sum of the estimated bandwidths used by ongoing rtPS and nrtPS connections ($\overline{E}_{\text{rtPS}}(t)$, $\overline{E}_{\text{nrtPS}}(t)$) and the estimated bandwidths to be used by the incoming connection (ΔN_{rtPS} or ΔN_{nrtPS}) is larger than a predefined upper threshold, the incoming connection is rejected; otherwise, the connection is accepted with certain probability depending on the estimated bandwidth usage and the connection priority. Specifically, when the estimated bandwidth usage is high or the priority of the incoming connection is low, the acceptance probability becomes small, and vice versa. A detailed description of the proposed CAC algorithm for rtPS connections is listed in pseudocode 2, where $N_{\text{th}}^{\max}$ and $N_{\text{th}}^{\min}$ are the maximum and minimum capacity threshold respectively, ρ_{rtPS} is a blocking probability that is used to differentiae priorities for different traffic types, and P_{rtPS} is the acceptance probability. The same algorithm is applied for nrtPS connections.

Algorithm 2 Connection admission control algorithm for rtPS connections

if $\overline{E}_{\text{rtPS}}(t) + \overline{E}_{\text{nrtPS}}(t) + \Delta N_{\text{rtPS}} > N_{\text{th}}^{\max}$ **then**
  Reject the incoming connection
else if $\overline{E}_{\text{rtPS}}(t) + \overline{E}_{\text{nrtPS}}(t) + \Delta N_{\text{rtPS}} < N_{\text{th}}^{\min}$ **then**
  Accept the incoming connection
else
  $P_{\text{rtPS}} = \rho_{\text{rtPS}} \cdot \frac{N_{\text{th}}^{\max} - (\overline{E}_{\text{rtPS}}(t) + \overline{E}_{\text{nrtPS}}(t) + \Delta N_{\text{rtPS}})}{N_{\text{th}}^{\max} - N_{\text{th}}^{\min}}$
end if

For BE connections, they are always accepted since they do not impose any QoS requirements.

IV. Simulation Results and Discussions

To evaluate the performance of the proposed downlink hierarchical resource management framework, a system-level simulation is performed in OPNET.

A. System Model

We consider the downlink of a single-cell IEEE 802.16 OFDMA/TDD system with cell radius of 2 km, where subscriber stations (SSs) are randomly placed in the cell with uniform distribution. The total bandwidth is set to be 5 MHz, which is divided into 10 subchannels. The duration of a frame is set to be 1 ms as recommended by the standard so that the channel quality of each connection almost remains constant within a frame, but may vary from frame to frame. We consider pass loss and large-scale shadowing as channel models. The modulation order and coding rate in AMC is determined by the instantaneous SNR.

Table I summarizes the system parameters used in the simulation. We assume that the base station (BS) has perfect knowledge of channel state information (CSI) of each subchannel of each user. We also assume that all packets are transmitted and received without errors and the transmission delay is negligible.

Parameters	Value
System	OFDMA/TDD, TDM
Central frequency	3500 MHz
Channel bandwidth	5 MHz
Number of subchannels	10
Length of OFDM symbol	156.25 μs
User distribution	Uniform
Beam pattern	Omni-directional
Cell radius	2 km
Frame duration	1 ms
BS transmit power	10 W
Pass loss model	Okumura-Hata model
Large-scale shadowing	Log-normal distribution with mean: 0, standard deviation: 8 dB
Maximum MAC PDU size	56 bytes

TABLE I

A SUMMARY OF SYSTEM PARAMETERS

Type	Characteristics	Distribution	Parameters
VoIP	ON period	Exponential	Mean = 1.34 sec
VoIP	OFF period	Exponential	Mean = 1.67 sec
VoIP	Packet size	Constant	66 bytes
VoIP	Inter-arrival time between packets	Constant	20 ms
Video	Packet size	Log-normal	Mean = 4.9 bytes Std. dev. = 0.75 bytes
Video	Inter-arrival time between packets	Normal	Mean = 33 ms Std. dev. = 10 ms
Web	Reading time between sessions	Exponential	Mean = 5 sec
Web	Number of packets within a packet call	Geometric	Mean = 25 packets
Web	Inter-arrival time between packets	Geometric	Mean = 0.0277 sec
Web	Packet size	Truncated Pareto	k = 81.5 bytes α = 1.1 m = 2 M bytes

TABLE II

A SUMMARY OF TRAFFIC PARAMETERS

B. Traffic Model

In the simulation, three traffic types are generated: VoIP, videoconference, and internet traffic. VoIP and videoconference are served in UGS class and rtPS class, respectively. Internet traffic is served in nrtPS class and BE class. Each user alternates between the states idle and busy, and generates one or several traffic types independently during the busy period. VoIP traffic is modeled as a two-state Markov ON/OFF source. A videoconference consists of a VoIP and a video source. Internet traffic can be web browsing that requires large bandwidth and generates variable size bursty data. We apply the WWW browsing model. A summary of traffic parameters for different traffic types are listed in Table II.

C. Performance Evaluation

Since the performance of fixed bandwidth allocation for UGS connections is well defined by the standard and BE connections do not have any specific QoS requirements, here we only focus on the performance evaluation of rtPS and nrtPS connections. The delay constraint for rtPS service is 50 ms and the minimum throughput constraint for nrtPS service is 100 Kbits/sec. The target outage probabilities for both rtPS and nrtPS services should be less than 3%.

The first two subplots in Fig. 2 & 3 show the estimated and the allocated bandwidths to rtPS and nrtPS class, respectively. It can be seen from the figure that the allocated bandwidth in rtPS class closely follows the pattern of the estimated bandwidth, while in nrtPS class, the allocated bandwidth does not always follow the pattern of the estimated bandwidth. This is due to the reason that the proposed resource allocation algorithm tend to allocate more bandwidth to nrtPS class when the estimated sum of bandwidth from rtPS and nrtPS classes is smaller than the total available bandwidth. In other words, the ARA first allocates E_{rtPS} and E_{nrtPS} to rtPS and nrtPS classes respectively, then most of the residual bandwidth is allocated to nrtPS class. The performance for rtPS class scheduler (packet delay and delay outage probability) and nrtPS class scheduler (minimum throughput outage probability) are depicted in the latter subplots of Fig. 2 & 3, respectively. One advantage of the proposed resource allocation algorithm is that instead of allocating the available bandwidth to each class scheduler in a static manner, the proposed algorithm adaptively allocates a "necessary" amount of bandwidth to each class scheduler to intentionally keep its outage probability around a predefined threshold, which is 3% in our scenario. By doing so, the channel and QoS aware class scheduler has more chances to serve a user in a good channel state without sacrificing the QoS requirement (as we can see in Fig. 2 & 3 that the packet delay in rtPS class is well kept below the maximum allowable delay and the outage probabilities in both classes fluctuates around the target threshold), thus significantly increase the efficiency of bandwidth utilization.

Fig. 4 shows the connection rejection probability in rtPS and nrtPS classes. It is obvious that the CAC decision for rtPS and nrtPS classes depends on the current state of the network (the estimated amount of bandwidth used by the ongoing connections). The acceptance probability of a new coming connection is inverse-proportional to the resource usage by the ongoing connections. Furthermore, we can see that the nrtPS connections have higher rejection probability than the rtPS connections. This is because rtPS class is given higher acceptance probability than nrtPS class.

To investigate the performance of the proposed resource management framework under different traffic loads, a series of simulations are performed with different number of users. The results are shown in Fig. 5. The average outage probabilities in both rtPS and nrtPS classes are well kept around the predetermined threshold regardless of the number of users, while the average connection rejection probabilities for both classes increases proportional to the number of users, thus preventing the system capacity from being overused.

V. CONCLUSIONS

This paper addresses the problem of radio resource management in OFDMA-based WiMAX systems. Specifically, we have proposed a hierarchical downlink resource management

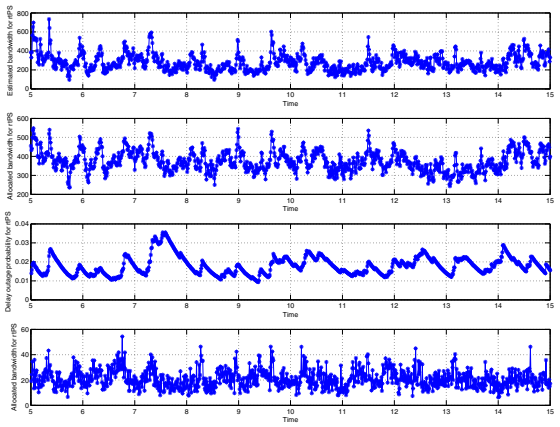

Fig. 2. Performance of rtPS class (estimated bandwidth, allocated bandwidth, packet delay and delay outage probability)

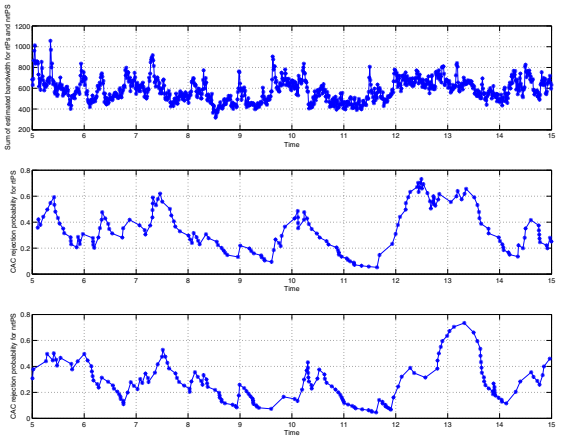

Fig. 4. CAC rejection probability of rtPS and nrtPS

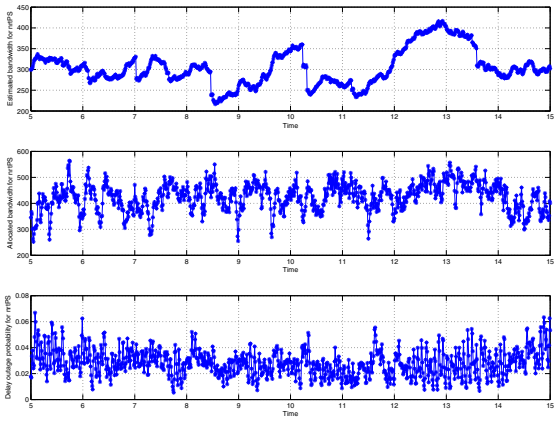

Fig. 3. Performance of nrtPS class (estimated bandwidth, allocated bandwidth, and minimum throughput outage probability)

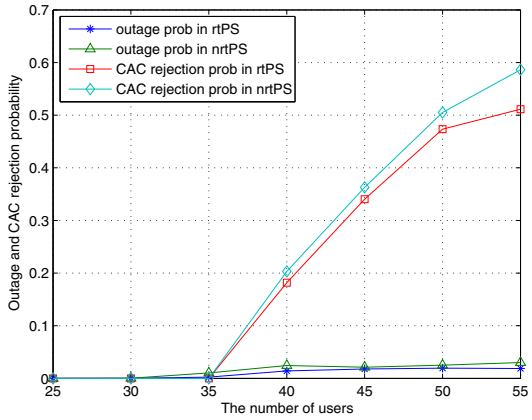

Fig. 5. Outage probability and CAC rejection probability

framework which consists of a DRA module and a CAC module. A priority-based scheduling algorithm for the class scheduler and an adaptive bandwidth allocation algorithm for the ARA are proposed for the DRA module. The advantage of this two-level hierarchical resource allocation architecture is that the scheduler and the allocator can be developed independently, yet there is still sufficient coupling between them as the allocator is aware of the performance of the scheduler. A measurement-based admission control strategy is proposed for the CAC module, which works cooperatively with the DRA module when admission decisions are made. Simulation results show that the proposed framework can work efficiently and adaptively to various kinds of traffic loads, and can satisfy the diverse QoS requirements in each service class.

REFERENCES

[1] Ying J. Z., and Khaled B.: *Energy-Efficient MAC-PHY Resource Management with Guaranteed QoS in Wireless OFDM Networks*, ICC 2005, Vol.5, pp. 3127–3131, May. 2005.

[2] Xing Z., and Wenbo W.: *Multiuser frequency-time domain radio resource allocation in downlink OFDM systems: Capacity analysis and scheduling methods*, Computers and Electrical Engineering, Vol.32 Issue.1-3, pp. 118–134, 2006.

[3] Kitti W. and Aura G.: *Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems*, International Journal of Communication Systems, Vol.16 Issue.1, pp. 81–96, 2003.

[4] Ali S.H., Ki-Dong Lee, and Leung V.C.M.: *Dynamic resource allocation in OFDMA wireless metropolitan area networks*, IEEE Wireless Communications, Vol.14 Issue.1, pp. 6–13, 2007.

[5] Rong Bo, Qian Yi, and Lu Kejie: *Integrated Downlink Resource Management for Multiservice WiMAX Networks*, IEEE Transactions on Mobile Computing, Vol.6 Issue.6, pp. 621–632, 2007.

[6] Majid G., and Raouf B.: *Call admission control in mobile cellular networks: a comprehensive survey*, Wireless Communications and Mobile Computing, Vol.6 Issue.1, pp. 69–93, 2006.

[7] Sanjay S., and Alexander L.S.: *Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR*, Proceedings of International Teletraffic Congress (ITC), 2001.

[8] Hua Wang: *Priority-based Resource Allocation for RT and NRT Traffics in OFDMA Systems*, The 3rd IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM), 2007.

[9] IEEE 802.16-2004, *IEEE standard for Local and Metropolitan Area Networks - Part 16: Air Interface for FBWA Systems*, 2004